Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Mária Špaková

Testování heteroskedasticity

Katedra pravěpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jan Kalina, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2011

Na tomto mieste by som sa rada poďakovala vedúcemu mojej bakalárskej práce RNDr. Janovi Kalinovi, Ph.D. za cenné rady a ochotu, s ktorou mi venoval svoj čas a za poskytnuté materiály, mojim rodičom a sestrám za všestrannú podporu počas celého môjho doterajšieho štúdia a môjmu priateľovi za trpezlivosť a podporu.

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, hlavne skutočnosť, že Univerzita Karlova v Prahe má právo na uzavretie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe dňa 15.4.2011 Mária Špaková

Obsah

Názov práce: Testování heteroskedasticity
Autor: Mária Špaková
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci bakalárskej práce: RNDr. Jan Kalina Ph.D., Ústav informatiky AV ČR

Abstrakt: Predložená práca sa zaoberá testovaním heteroskedasticity. Je rozdelená na štyri kapitoly, pričom prvé tri sa zaoberajú teóriou a posledná je venovaná praktickému testovaniu na konkrétnych dátach.

V úvode teoretickej časti sú zhrnuté základné pojmy, poznatky a vzťahy týkajúce sa lineárnej regresie, regresného modelu a odhadu parametrov metódou najmenších štvorcov. Nasleduje časť venovaná heteroskedasticite, jej dôsledkom a riešeniu, spolu s popisom testov heteroskedasticity: Breuschov - Paganov, Goldfeldov - Quandtov a Whiteov.

V praktickej časti práce sú aplikované popisované testy spolu s inými metódami na odhalenie heteroskedasticity na dáta v troch príkladoch: Výdaje vs. príjmy, HDP a Výdavky na potraviny.

Cieľom práce je diskutovať o vyššie spomenutých testoch. V konkrétnych príkladoch sa okrem iného ukazuje, že testovanie heteroskedasticity môže viesť k rôznym výsledkom pre tie isté dáta, čo potvrdzujú i vzťahy uvedené v teoretickej časti. Pretože neexistuje optimálny test, který by mal za každých okolností lepšie vlastnosti než všetky ostatné testy, nie je možné jednoznačne doporučiť len jeden konkrétny test pre všetky konkrétne aplikácie.

Kľúčové slová: lineárna regresia, heteroskedasticita, Goldfeldov - Quandtov test, Breuschov - Paganov test, Whiteov test

Title: Testing heteroscedasticity
Author: Mária Špaková
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Jan Kalina Ph.D., Institute of Computer Science, Academy of Sciences of the Czech Republic

Abstract: This paper deals with testing heteroscedasticity. It is divided into four chapters. The first three chapters focus on the theory and the last one is devoted to practical testing using specific data.

In the beginning of the theoretical part, basic concepts, knowledge and relationships concerning the linear regression, the regression model and the estimation of parameters by the method of ordinary least squares are introduced. The rest

of this part is devoted to heteroskedasticity, its consequences and solutions. The following heteroscedasticity tests are being discussed: Breusch - Pagan, Goldfeld - Quandt and White.

The practical part contains actual applications of the described tests and other methods to detect heteroskedasticity using three examples: Outlays vs. income, GDP and Expenditures on food.

The aim of this paper is to discuss the above-mentioned tests. Three examples on real data with economic motivation confirm the theoretical properties of the tests. A uniformly optimal test of heteroscedasticity does not exist and different tests yield rather different results.

Keywords: linear regression, heteroscedasticity, Goldfeld - Quandt test, Breusch - Pagan test, White test

Úvod

Heteroskedasticita je jav, s ktorým sa bežne stretávame, no mnohokrát mu nevenujeme dostatočnú pozornosť. Veľmi často sa vyskytuje v makroekonomických regresných modeloch, ale môžeme na ňu natrafiť i pri spracovávaní modelov z iných oblastí života. Je to prípad, keď rozptyl náhodných chýb v lineárnom regresnom modeli nie je pre všetky pozorovania rovnaký, teda je porušený jeden z predpokladov pre odhadovanie regresných parametrov metódou najmenších štvorcov.

V prípade ignorovania heteroskedasticity môžeme pri svojich výskumoch dôjsť k nepresným alebo chybným výsledkom, čo nám môže spôsobiť nemalé problémy. Za prítomnosti heteroskedasticity sa totiž napríklad zvyšuje významnosť menej významných parametrov. Takisto nie je možné používať koeficient determinácie, konfidenčné intervaly alebo testy hypotéz o regresných parametroch a pod.

Keďže heteroskedasticita je závažným problémom, je dôležité ju vedieť testovať. Práve jej testovanie je jadrom tejto práce. Okrem klasických grafických metód odhaľovania heteroskedasticity existuje mnoho testov, ktoré nám môžu pomôcť zistiť jej prítomnosť. S niektorými z nich sa zoznámime v tejto práci, a to nielen teoreticky, ale i prakticky.

Zatiaľ čo heteroskedasticita sa môže vyskytnúť v ľubovoľných regresných modeloch, táto práca se venuje iba téme lineárna regresia. Prvá kapitola je preto venovaná práve jej. Zhrnieme v nej dôležité fakty, vety a vzťahy, ktoré súvisia s lineárnou regresiou, regresným modelom a metódou najmenších štvorcov. Druhá kapitola je venovaná heteroskedasticite, jej dôsledkom, detekcii a heteroskedastickej regresii. Na ňu nadväzuje tretia kapitola, v ktorej sa zoznámime s najbežnejšími testami používajúcimi sa v praxi na testovanie heteroskedasticity. Tieto testy následne vo štvrtej kapitole aplikujeme na konkrétne dáta.

Cieľom tejto práce je poukázať na dôležitosť testovania heteroskedasticity a na fakt, že je potrebné pristupovať k nej citlivo. Jednotlivé testy, ktoré budú použité, je potrebné voliť opatrne vzhľadom k povahe heteroskedasticity. Nie je ich vhodné používať úplne všeobecne, pretože majú svoje obmedzenia a niektoré z nich sú navyše ovplyvnené subjektívnymi voľbami. V praxi se testy heteroskedasticity používajú úplne automaticky. V tejto práci však upozorníme na fakt, že jednotlivé testy majú svoje konkrétne predpoklady, ktoré by mali byť splnené.

Kapitola 1

Lineárna regresia

Prvá kapitola tejto práce sa zaoberá základnými pojmami týkajúcimi sa lineárnej regresie. V podkapitole 1.1 je popísaný klasický regresný model a sú v nej spomenuté dôležité vety a vzťahy, ktoré s klasickým regresným modelom súvisia. Táto podkapitola sa taktiež venuje odhadu parametrov metódou najmenších štvorcov. Podkapitola 1.2 je venovaná špeciálne lineárnej regresii s jedným regresorom. Väčšina teoretických poznatkov nachádzajúcich sa v tejto kapitole pochádza z Anděla [1] alebo Cipru [3].

1.1 Klasický regresný model

Majme náhodné veličiny $Y_1, ..., Y_n$ a maticu daných čísel $\mathbf{X} = (X_{ij})$ typu $n \times k$, kde $k < n$. Predpokladajme, že pre náhodný vektor $\mathbf{Y} = (Y_1, ..., Y_n)'$ platí

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \tag{1.1}$$

kde $\beta = (\beta_1, ..., \beta_k)'$ je vektor neznámych parametrov a $\mathbf{e} = (e_1, ..., e_n)'$ je náhodný vektor spĺňajúci podmienky: $\mathsf{E}(e) = \mathbf{0}$, $\mathsf{var}(e) = \sigma^2\mathbf{I}$. Pritom $\sigma^2 > 0$ je taktiež neznámy parameter. Tento model budeme nazývať *regresným modelom*. Pretože $\mathbf{Y}$ závisí na β lineárne, hovoríme o *lineárnom regresnom modeli* alebo jednoducho o *lineárnej regresii*. Je potrebné si uvedomiť, že vektor $\mathbf{X}\beta$ je nenáhodný. Preto platí: $\mathsf{E}(\mathbf{Y}) = \mathbf{X}\beta$ a $\mathsf{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Parametre $\beta_1, ..., \beta_k$ sa odhadujú *metódou najmenších štvorcov*, tj. z podmienky, že výraz $(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$ ako funkcia β má byť minimálny. Tieto odhady označíme $\mathbf{b} = (b_1, ..., b_k)'$.

Často sa stáva, že prvý stĺpec matice $\mathbf{X}$ je tvorený iba jednotkami. Položme $p = k - 1$. Potom môžeme písať: $\mathbf{X} = (\mathbf{1}, \mathbf{X_1})$, kde $\mathbf{X_1}$ je matica typu $n \times p$. Vektory β a $\mathbf{b}$ sa potom zapisujú v tvare: $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ a $\mathbf{b} = (b_0, b_1, ..., b_p)'$.

V nasledujúcom texte budeme používať niekoľko dôležitých viet, ich dôkazy však neuvádzame, je možné ich nájsť v literatúre [1].

Veta 1 *Odhady metódou najmenších štvorcov sú:* $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Sústava lineárnych rovníc $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ sa nazýva *sústava normálnych rovníc*, resp. *normálne rovnice*. Vektor $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ sa nazýva *vypočítaná hodnota metódou najmenších štvorcov* a môže byť považovaný za najlepšiu aproximáciu vektoru $\mathbf{Y}$, aká sa dá vytvoriť lineárnou kombináciou stĺpcov matice $\mathbf{X}$. Označme $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ a $\mathbf{M} = \mathbf{I} - \mathbf{H}$. Je vidieť, že $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. Označenie $\mathbf{H}$ vzniklo z anglického termínu *hat matrix.* Ide o projekčnú maticu a projekcia sa často označuje strieškou nad príslušným symbolom. Je zrejmé, že $\mathbf{H}$ je symetrická a idempotentná. Preto je symetrická a idempotentná i matica $\mathbf{M}$. Pritom máme

$$h(\mathbf{H}) = Tr\mathbf{H} = Tr\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = Tr\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = Tr\mathbf{I}_k = k, \tag{1.2}$$

takže $h(\mathbf{M}) = Tr\mathbf{M} = n - k$.

Veta 2 *Platí:* $\mathsf{E}(\mathbf{b}) = \beta$, $\mathsf{var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

To znamená, že odhad metódou najmenších štvorcov $\mathbf{b}$ je nestranným odhadom vektoru parametrov β.

Veličinu

$$\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} \tag{1.3}$$

nazývame reziduá, teda odhad nepozorovateľných hodnôt reziduálnej zložky, resp. disturbancie $\mathbf{e}$ a veličine

$$SSE = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \tag{1.4}$$

hovoríme *rezíduálny súčet štvorcov (error sum of squares).*

Veta 3 *Pre rezíduálny súčet štvorcov platí:*

$$\text{SSE} = \mathbf{Y}'\mathbf{M}\mathbf{Y} \; a \; \text{SSE} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}.$$

Teraz si popíšeme vlastnosti odhadu metódou najmenších štvorcov. Ich odvodenie je však možné len v prípade, že model spĺňa určité predpoklady. Predpoklady charakterizujúce klasický model lineárnej regresie (1.1) sa často uvádzajú v nasledujúcom tvare:

- $\mathsf{E}(e_i) = 0$, tj. stredná hodnota rezíduálnej zložky je nulová pre všetky i;

- $\mathsf{var}(e_i) = \sigma^2 < \infty$, tj. rozptyl rezíduálnej zložky je konštantný a konečný pre všetky i;
- $\mathsf{cov}(e_s, e_t) = 0$ pre $s \neq t$, tj. rezíduálne zložky sú navzájom nekorelované pre všetky $s \neq t$;
- $\mathsf{cov}(X_{ij}, e_t) = 0$, tj. regresory sú v rovnakom čase alebo pre rovnakú priezrezovú jednotku nekorelované s rezíduálnou zložkou pre všetky i a j;
- $h(\mathbf{X}) = k$, tj. náhodná matica $\mathbf{X}$ má lineárne nezávislé stĺpce.

V prípade splnenia týchto predpokladov má odhad metódou najmenších štvorcov tieto vlastnosti:

- je *nestranný*, teda jeho stredná hodnota je rovná hodnote odhadovaného parametru,
- je *konzistentný*, teda pri rastúcom rozsahu výberu n konverguje v pravdepodobnosti ku skutočnej hodnote odhadovaného parametru.

Detailnejšie vysvetlené a popísané predpoklady, rovnako ako aj vlastnosti odhadu metódou najmenších štvorcov, je možné nájsť v napríklad u Cipru [3].

Veta 4 *Náhodná veličina*

$$s^2 = \frac{SSE}{(n-k)}$$

je nestranným odhadom parametru σ^2.

Označme

$$SST = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 \tag{1.5}$$

Veličinu SST nazývame *celkovým súčtom štvorcov (total sum of squares)*. Pre úplnosť spomenieme i vzťah

$$SSR = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}), \tag{1.6}$$

ktorý vyjadruje *regresný súčet štvorcov (regression sum of squares)*. Pre spomínané súčty štvorcov platí, ako uvádza Kmenta [12],

$$SST = SSR + SSE, \tag{1.7}$$

resp.

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}. \tag{1.8}$$

Všetky tieto súčty, teda *SSE*, *SSR* a *SST* sú prehľadne uvedené i v nasledujúcej tabuľke:

Tabuľka 1.1: Súčty štvorcov

SSE	$\sum(Y_i - \hat{Y}_i)^2$
SST	$\sum(Y_i - \bar{Y})^2$
SSR	$\sum(\hat{Y}_i - \bar{Y})^2$

K popisu presnosti regresného modelu sa používa *koeficient determinácie* R^2 a *korigovaný koeficient determinácie* $\bar{R}^2$, ktoré sú dané vzorcami:

$$R^2 = 1 - \frac{SSE}{SST}, \tag{1.9}$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-r-1} \cdot \frac{SSE}{SST}. \tag{1.10}$$

Čím sú tieto koeficienty bližšie jednej, tým tesnejšia je lineárna regresná závislosť. Pri aplikáciách regresnej analýzy sa niekedy stretávame s predpokladom normality. Všetky doterajšie tvrdenia platili bez použitia tohto predpokladu, ďalej ho však už budeme potrebovať. Budeme teda predpokladať, že $\mathbf{e} \sim \mathrm{N}(\mathbf{0},\sigma^2\mathbf{I})$, z čoho plynie, že $\mathbf{Y} \sim \mathrm{N}(\beta,\sigma^2(\mathbf{X'X})^{-\mathbf{1}})$.

Veta 5 *Platí:* $\mathbf{b} \sim N(\beta, \sigma^2(\mathbf{X'X})^{-\mathbf{1}})$.

Veta 6 *Platí:* $\frac{SSE}{\sigma^2} \sim {\chi^2}_{n-k}$.

Veta 7 *Vektor* $\mathbf{b}$ *a veličina* s^2 *sú nezávislé.*

Vo zvyšku tejto kapitoly budeme popisovať testy hypotéz a konfidenčné intervaly pre β.

Veta 8 *Označme* v_{ij} *prvky matice* $(\mathbf{X'X})^{-1}$. *Nech*

$$T_i = \frac{(b_i - \beta_i)}{\sqrt{s^2 v_{ij}}}.$$

Potom pre každé i = 1, ..., k platí $T_i \sim t_{n-k}$.

Pomocou tejto vety je možné testovať hypotézu $H_0 : \beta_i = \beta_i^0$, kde i je nejaké pevne dané číslo. Napríklad pri alternatíve $H_1 : \beta_i \neq \beta_i^0$ zamietneme H_0 na hladine α, ak platí:

$$\frac{|b_i - \beta_i^0|}{\sqrt{s^2 v_{ii}}} \geq t_{n-k}(\alpha) \tag{1.11}$$

Najčastejším prípadom je $\beta_i{}^0 = 0$. Potom sa overuje, či $\mathbf{Y}$ vôbec závisí na i-tom stĺpci matice $\mathbf{X}$ alebo či je možné vypustením i-tého stĺpca prejsť k jednoduchšiemu modelu. Častokrát je však potrebné testovať hypotézu s niekoľkými parametrami naraz. Položme

$$\beta = \binom{\beta_1}{\beta_2}\ ,\ \mathbf{b} = \binom{b_1}{b_2}\ ,\ (\mathbf{X'X})^{-\mathbf{1}} = \begin{pmatrix} \mathbf{V} & \mathbf{U} \\ \mathbf{U'} & \mathbf{W} \end{pmatrix},$$

kde β_1 a $\mathbf{b}_1$ majú p zložiek (tu se však p používa v inom zmysle ako sme uviedli na začiatku tejto kapitoly), β_2 a $\mathbf{b}_2$ majú q zložiek (pričom $p + q = k$), $\mathbf{V}$ je matica typu $p \times p$ a $\mathbf{W}$ je typu $q \times q$.

Veta 9 *Platí:* $Z = \frac{1}{qs^2}\ (\mathrm{b}_2 - \beta_2)'\mathbf{W}^{-1}(\mathrm{b}_2 - \beta_2) \sim F_{q,n-k}$.

Táto veta sa používa na testovanie hypotézy $H_0 : \beta_2 = \beta_2^0$ proti alternatíve $H_1 : \beta_2 \neq \beta_2^0$. V prípade, keď

$$Z = \frac{1}{qs^2}(b_2 - \beta_2)'\mathbf{W}^{-1}(b_2 - \beta_2) \geq F_{q,n-k}(\alpha), \tag{1.12}$$

zamietame H_0 na hladine α. Najčastejšie býva β_2^0 rovné nule. Tým sa testuje hypotéza, že vektor $\mathbf{Y}$ nezávisí na posledných q stĺpcoch matice $\mathbf{X}$. Veta platí aj pre $q = k$, ak položíme $\mathbf{W} = (\mathbf{X'X})^{-1}$. Z tejto vety vyplýva, že v prípade testu $H_0 : \beta = \mathbf{0}$ proti $H_1 : \beta \neq \mathbf{0}$ platí

$$Z = \frac{(\mathbf{Y'Y}) - SSE}{ks^2}. \tag{1.13}$$

Ak platí H_0, máme $Z \sim \mathrm{F}_{k,n-k}$.

Teraz ukážeme, ako sa postupuje v prípade, keď sa musíme zaoberať nejakou danou lineárnou kombináciou zložiek vektoru β.

Veta 10 *Nech* $\mathbf{c} = (c_1, ..., c_k)'$ *je daný nenulový vektor. Potom* $E\mathbf{c'b} = \mathbf{c'}\beta$ *a*

$$T = \frac{\mathbf{c'b} - \mathbf{c'}\beta}{\sqrt{s^2\mathbf{c'}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{c}}} \sim t_{n-k}.$$

Z tejto vety vyplýva, že obojstranný interval spoľahlivosti $\mathbf{c'}\beta$ s koeficientom spoľahlivosti 1 - α je:

$$(\mathbf{c'b} - t_{n-k}(\alpha)\sqrt{s^2\mathbf{c'}(\mathbf{X'X})^{-1}\mathbf{c}}, \mathbf{c'b} + t_{n-k}(\alpha)\sqrt{s^2\mathbf{c'}(\mathbf{X'X})^{-1}\mathbf{c}}). \tag{1.14}$$

1.2 Lineárna regresia s jedným regresorom

Lineárny regresný model s jedným regresorom je najjednoduchším regresným modelom. Je však podobne ako modely s viacerými regresormi citlivý voči heteroskedasticite. Dokonca i samotná minimalizácia súčtu štvorcov reziduí vo svojej podstate vychádza z predpokladu, že každé pozorovanie má chybu s rovnakým rozptylom.

Nech n $\geq$ 3 a nech platí

$$Y_i = \beta_0 + \beta_1 X_i + e_i,\ i = 1, ..., n. \tag{1.15}$$

Máme

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ ... & ... \\ 1 & X_n \end{pmatrix}, \mathbf{X'X} = \begin{pmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{pmatrix}, \mathbf{X'Y} = \begin{pmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{pmatrix}.$$

Matica $\mathbf{X}$ bude mať plnú hodnosť v prípade, že vektor $(X_1, ..., X_n)'$ nebude obsahovať iba rovnaké prvky. Označme

$$\bar{X} = \frac{\sum X_i}{n}, \tag{1.16}$$

$$\bar{Y} = \frac{\sum Y_i}{n}, \tag{1.17}$$

kde $i = 1, ..., n$. Odhady metódou najmenších štvorcov sú:

$$b_0 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}, \tag{1.18}$$

$$b_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}. \tag{1.19}$$

Podľa týchto vzorcov sa zvyčajne počíta len b_1, zatiaľ čo b_0 sa vypočíta zo vzťahu

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{1.20}$$

ktorý plynie z prvej rovnice sústavy normálnych rovníc. Ďalej máme

$$s^2 = \frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i}{n - 2} \tag{1.21}$$

Najčastejšie sa zaoberáme parametrom β_1. Test H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$ zakladáme na štatistike $T_1 = b_1 \sqrt{\sum X_i^2 - n\bar{X}^2}/s$. V prípade $|T_1| \geq t_{n-2}(\alpha)$ hypotézu H_0 zamietame. Interval spoľahlivosti pre β_1 s koeficientom spoľahlivosti 1 - α je

$$(b_1 - \frac{t_{n-2}(\alpha)s}{\sqrt{\sum X_i^2 - n\bar{X}^2}}, b_1 + \frac{t_{n-2}(\alpha)s}{\sqrt{\sum X_i^2 - n\bar{X}^2}}). \tag{1.22}$$

Často sa zaujímame o hodnotu $\beta_0 + \beta_1 X$, kde X je nejaké dané číslo. Používame pritom Vetu 10. Máme $\mathbf{c} = (1, X)'$. Nestranným odhadom pre $\beta_0 + \beta_1 X$ *je* $b_0 + b_1 X$. Máme

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum X_i^2 - n\bar{X}^2} \begin{pmatrix} \frac{\Sigma X_i^2}{n} & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}, \tag{1.23}$$

$$c'(\mathbf{X}'\mathbf{X})^{-1}c = \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X_i^2 - n\bar{X}^2}. \tag{1.24}$$

Interval spoľahlivosti pre $\beta_0 + \beta_1 X$ s koeficientom spoľahlivosti 1 - α je preto interval s koncovými bodmi

$$b_0 + b_1 X \pm t_{n-2}(\alpha)s\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum X_i^2 - n\bar{X}^2}}. \tag{1.25}$$

Obvykle nás zaujímajú hodnoty X z určitého intervalu. Pre ne sa potom graficky znázorní ako $b_0 + b_1 X$, tak i medze vyššie uvedeného intervalu spoľahlivosti. Tieto medze pri spojite sa meniacom X vytvoria tzv. *pás spoľahlivosti okolo regresnej priamky.* Je zrejmé, že tento pás má najmenšiu šírku v bode $X = \bar{X}$. S rastúcou vzdialenosťou od $\bar{X}$ šírka pásu monotónne rastie. Nie je ho však možné interpretovať tak, že s pravdepodobnosťou 1 - α pokrýva celú teoretickú regresnú priamku $b_0 + b_1 X$.

Kapitola 2

Heteroskedasticita

Jedným z predpokladov charakterizujúcim tzv. klasický model lineárnej regresie, je konštantný rozptyl rezíduálnych zložiek. Tento predpoklad sa nazýva *homoskedasticita* a uvádza sa v nasledujúcom tvare

$$\mathsf{var}(\mathbf{e}) = \sigma^2 < \infty. \tag{2.1}$$

V prípade porušenia homoskedasticity hovoríme o *heteroskedasticite*. Teda ak rezíduálne zložky nemajú konštantný rozptyl, potom sa označujú ako *heteroskedastické*.

Predtým, ako sa budeme hlbšie zaoberať heteroskedasticitou, zavedieme pojem *zobecnený model lineárnej regresie*, ktorý používa Cipra [3]. Je to model, ktorý predpoklad konštantného rozptylu rezíduálnych zložiek a ich nekorelovanosti zobecňuje do tvaru

$$\mathsf{var}(\mathbf{e}) = \sigma^2\Omega, \tag{2.2}$$

kde $\mathbf{\Omega}$ je pozitívne definitná matica, tj. rozptyl rezíduálnej zložky nemusí byť konštantný a rezíduálne zložky nemusia byť navzájom nekorelované. Inak platia ostatné predpoklady klasického modelu lineárnej regresie, o ktorých sme hovorili v predchádzajúcej kapitole, teda

- $\mathsf{E}(e_i) = 0$;
- $\mathsf{var}(X_{ij}, e_i) = 0$;
- $h(\mathbf{X}) = k$;
- príp. $e_i \sim N(0, \sigma^2)$.

V rámci zobecneného modelu lineárnej regresie predstavuje heteroskedasticita prípad, keď

$$\mathsf{var}(\mathbf{e}) = \sigma^2\Omega = \sigma^2 \cdot \mathbf{diag}\{k_1, ..., k_n\}, \sigma^2 > 0, k_1, ..., k_n > 0, \tag{2.3}$$

t. j. rezíduálne zložky e_i majú nekonštantný rozptyl $\sigma^2 k_i$ s neznámymi kladnými hodnotami k_i a sú navzájom nekorelované.

2.1 Dôsledky heteroskedasticity

Dôsledkami heteroskedasticity rozumieme dôsledky, ktoré sa prejavia v prípade, keď ignorujeme heteroskedasticitu modelu a použijeme klasický odhad metódou najmenších štvorcov. Tieto dôsledky môžu byť závažné, ak je homoskedasticita výrazne porušená. Medzi dôsledky heteroskedasticity patria nasledujúce prípady:

- odhad **b** metódou najmenších štvorcov zostáva nestranným a konzistentným odhadom parametrov β, ale nie je všeobecne najlepším odhadom medzi nestrannými lineárnymi odhadmi parametrov β ;
- odhad s^2 metódou najmenších štvorcov nie je všeobecne nestranným odhadom parametru σ^2;
- nie je možné používať konfidenčné intervaly a testy hypotéz o regresných parametroch β ;
- nie je možné používať hodnotu koeficientu determinácie R^2;
- zvyšuje sa významnosť menej významných parametrov, pretože hodnota testovej štatistiky je väčšia ako by bola v prípade homoskedasticity.

Overenie predpokladov zhody rozptylov chýb $\mathbf{e} = (e_1, ..., e_n)'$ je možné previesť rôznymi spôsobmi, a to na základe rezíduí $\mathbf{u} = (u_1, ..., u_n)'$, kde

$$u_i = Y_i - b_1 X_{1i} - ... - b_k X_{ki},\ i = 1, ..., n, \tag{2.4}$$

pretože práve rezíduá môžeme vnímať ako odhady nepozorovaných chýb **e**.

2.2 Detekcia heteroskedasticity

Odhaliť heteroskedasticitu môžeme na základe hrubších a subjektívnych kritérií, napr.:

- graf rezíduí oproti vyhladeným hodnotám $\hat{\mathbf{Y}} = \mathbf{Xb}$;

- graf rezíduí oproti jednému z regresorov;
- graf rezíduí $(1, u_1)$, $(2, u_2)$, ..., (n, u_n), tzv. index plot rezíduí.

Tieto kritéria môžu poskytnúť nápad, akým spôsobom heteroskedasticitu popísať a modelovať, čo je prínosné nielen pre jej testovanie pomocou testu hypotéz, ale i následné odstránenie.

Rozpoznanie heteroskedasticity pomocou grafu je však použiteľné, iba ak poznáme príčinu heteroskedasticity, a to v praxi nebýva časté. Pôvodné testy heteroskedasticity sa týkali situácií s rozpoznateľnou príčinou heteroskedasticity, keď je napríklad možné usporiadať pozorovanie podľa veľkosti regresoru, ktorý lineárne ovplyvňuje smerodajnú odchýlku chýb. Existuje aj mnoho formálnych testov heteroskedasticity, ktoré nevyžadujú príliš veľa predbežných informácií o pravdepodobnom tvare heteroskedasticity. Niektorým z týchto testov sa budeme venovať v tretej kapitole.

2.3 Heteroskedastická regresia

Ak zamietame nulovú hypotézu zhody rozptylov chýb v lineárnom modeli (1.1), mal by sa tento model transformovať na iný model tak, aby sa odstránili nežiaduce dôsledky heteroskedasticity. Odhad regresných parametrov v transformovanom modeli potom nazývame *heteroskedastická regresia*.

Ak máme určitú predstavu o forme heteroskedasticity tak, ako tomu bude ďalej pri prevedení Goldfeldovho - Quandtovho testu, teda

$$H_1 : \mathsf{var}(\mathsf{e}) = \sigma^2 \cdot \mathbf{diag}\{k_1, ..., k_n\}, \tag{2.5}$$

podľa Greena [7] využijeme túto predstavu pre odstránenie heteroskedasticity. Pracujeme teda s modelom

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\beta_1 X_{1i}}{\sqrt{k_i}} + ... + \frac{\beta_p X_{pi}}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}}, \ i = 1, ..., n. \tag{2.6}$$

Jedným z typických príkladov je voľba $\sqrt{k_i} = X_{ji}$, kde $i = 1, ..., n$ a rozptyl chýb berieme priamoúmerný j-tému regresoru. Iné príklady zahŕňajú $\sqrt{k_i} = \sqrt{X_{ji}}$ alebo $\sqrt{k_i} = \hat{Y}_i = b_1 X_{1i} + ... + b_p X_{pi}$, $i = 1, ..., n$.

V modeli (2.6) odhadneme regresné parametre metódou najmenších štvorcov a doporučuje sa, aby sme znova testovali heteroskedasticitu. Ak sa tentokrát nezamieta nulová hypotéza homoskedasticity, považujeme tento model za vhodnejší než pôvodný. Preto výsledky berieme iba z transformovaného modelu, a to nielen

bodové odhady β, ale aj konfidenčné intervaly a testy hypotéz pre β, hodnotu koeficientu determinácie R^2 i ďalších štatistík.

Niekedy sa variabilita chýb modeluje zložitejším spôsobom, tak ako vo vzorci

$$\mathsf{var}(e_i) = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_K Z_{Ki},\ i = 1, ..., n \tag{2.7}$$

pri prevedení Breuschovho - Paganovho testu, ktorému sa venujeme v samostatnej kapitole. V tomto prípade prebieha odstránenie heteroskedasticity v dvoch krokoch. V prvom kroku najprv odhadneme regresné parametre v pôvodnom lineárnom modeli (1.1) metódou najmenších štvorcov a spočítame

$$u_i^2 = (Y_i - b_1 X_{1i} - ... - b_p X_{pi})^2 \tag{2.8}$$

Potom odhadneme regresné parametre v pomocnom regresnom modeli

$$u_i^2 = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_K Z_{Ki} + v_i,\ i = 1, ..., n, \tag{2.9}$$

kde $v_1, ..., v_n$ sú náhodné chyby. Tak získame odhady $\hat{\alpha}_0, ..., \hat{\alpha}_K$ pre regresné parametre $\alpha_0, ..., \alpha_K$. V druhom kroku použijeme vyhladené hodnoty u_i^2, čo sú hodnoty

$$\hat{u_i^2} = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{1i} + ... + \hat{\alpha}_K Z_{Ki},\ i = 1, ..., n, \tag{2.10}$$

ako konštanty $k_1, ..., k_n$ pre transformovaný model (2.6).

Riešenie heteroskedasticity je jednoduché, ak poznáme jej príčiny a tieto príčiny je naviac možné modelovo zvládnuť. Ukážeme si to na príklade prevzatom z Cipru [3]. Nech má heteroskedasticita napr. tvar

$$\mathsf{var}(e_i) = \sigma^2 z_i^2,\ i = 1, ..., n, \tag{2.11}$$

kde z_i sú pozorované hodnoty. Niekedy je možné vziať za z_i priamo hodnoty jedného z regresorov modelu. Uvažujme napr. *model úspor domácnosti*, v ktorom sú úspory S_i jednotlivých domácnosti modelované pomocou lineárnej regresie v závislosti na ich príjme I_i a ďalších faktoroch, ktoré zhrnieme schematicky pod symbol F_i (napr. súčasná spotreba, oneskorená spotreba a i.)

$$S_i = \beta_1 + \beta_2\ I_i + F_i + e_i. \tag{2.12}$$

Väčšinou ide o silno heteroskedastický model, v ktorom sa priamo ponúka zvoliť $\mathrm{z}_i = \mathrm{I}_i$, lebo potom

$$\sigma(S_i) = \sqrt{\mathsf{var}(S_i)} = \sigma(e_i) = \sigma z_i = \sigma I_i,\ i = 1, ..., n \tag{2.13}$$

má interpretáciu zodpovedajúcu ekonomickej realite: variabilita úspor u domácností s veľkými prijímami totiž býva oveľa väčšia ako u domácností s malými prijímami. Pozitívna lineárna závislosť smerodajnej odchýlky úspor na veľkosti príjmu je teda kompatibilná s praxou.

Pôvodný model je vhodné transformovať do tvaru

$$\frac{Y_i}{z_i} = \beta_1 \frac{1}{z_i} + \beta_2 \frac{X_{i2}}{z_i} + ... + \beta_k \frac{X_{ik}}{z_i} + u_i, \ i = 1, ..., n \tag{2.14}$$

kde teraz

$$\mathsf{var}\,(u_i) = \mathsf{var}\left(\frac{e_i}{z_i}\right) = \sigma^2, \ i = 1, ..., n, \tag{2.15}$$

tj. transformovaný model je homoskedastický, a teda optimálne odhadnuteľný pomocou klasického odhadu metódou najmenších štvorcov.

V praxi však obvykle príčiny heteroskedasticity nepoznáme. Napriek tomu, že existujú veľmi sofistikované procedúry, ktoré ponúkajú teoreticky podložené riešenia i v takomto prípade, vzhľadom na numerickú náročnosť takýchto metód a absenciu príslušného software sa v praxi dáva prednosť jednoduchším postupom. Jednou z možností je napr. aplikácia logaritmickej či inej transformácie na premenné tak, aby došlo k redukcii ich veľkosti vrátane redukcie prípadných extrémnych hodnôt, ktoré môžu heteroskedasticitu spôsobiť.

Iný postup pre ošetrenie heteroskedasticity popísal Cragg [4]. Jeho návrh je dvojstupňová metóda pre odhad regresných parametrov, ktorú využíva priamo pre odhad regresných parametrov pomocnej premennej. Odporúča sa ich voliť tak, aby prispievali k vysvetleniu variability odozvy. Robustnú obdobu Craggovho prístupu, ktorá nie je citlivá voči predpokladu normálneho rozdelenia chýb ani k prítomnosti odľahlých pozorovaní, navrhol Kalina [11].

Kapitola 3

Testy heteroskedasticity

Heteroskedasticita predstavuje potenciálne vážne problémy pri záveroch založených na metóde najmenších štvorcov. Málokedy si môžeme byť istí, že dáta sú heteroskedastické a ak sú, o akú formu heteroskedasticity ide. Je preto veľmi užitočné, aby sme boli schopní testovať homoskedasticitu a ak je to nutné, modifikovať podľa toho procedúry odhadov.

Väčšina testov heteroskedasticity je podľa Greena [7] založená na nasledujúcej stratégii: štandardná metóda najmenších štvorcov dáva zhodný odhad β aj za prítomnosti heteroskedasticity. Štandardné rezíduá potom napodobňujú, i keď nedokonalo kvôli výberovému rozptylu, heteroskedasticitu skutočných disturbancií. Preto sa testy vytvorené na detekciu heteroskedasticity vo väčšine prípadov aplikujú na rezíduá z najmenších štvorcov.

Testom heteroskedasticity sa rozumie test nulovej hypotézy:

$$H_0 : \mathsf{var}(e_i) = \sigma^2,\ i = 1, ..., n, \tag{3.1}$$

buď proti alternatívnej hypotéze:

$$H_1 : \mathsf{var}(e_i) \neq \sigma^2,\ i = 1, ..., n, \tag{3.2}$$

alebo proti alternatívnej hypotéze, ktorá je špeciálnym prípadom H_1. Takúto alternatívnu hypotézu považujeme za konkrétny model, ktorý popisuje štruktúru heteroskedasticity. Je žiadúce, aby alternatívna hypotéza čo najviac odpovedala skutočnej štruktúre heteroskedasticity.

V literatúre je popísané množstvo testovacích procedúr, napr. *Goldfeldov - Quandtov test*, *Breuschov - Paganov test* či *Whiteov test*, s ktorými sa zoznámime v tejto kapitole. Medzi ďalšie používané testy patrí napríklad *Bartlettov test*, ktorému sa venuje Zvára [16] alebo Szroeterova trieda testov, ktorou sa zaoberá Víšek [14]. Analogické testy heteroskedasticity pre robustnú regresiu odvodil Kalina [10].

3.1 Goldfeldov - Quandtov test

Goldfeldov - Quandtov test navrhli v roku 1965 Stephen M. Goldfeld a Richard E. Quandt [6]. Tento test je populárny a dá sa ľahko spočítať a interpretovať. Pri jeho použití sa vyžaduje alternatívna hypotéza v tvare:

$$H_1 : \mathsf{var}(\mathbf{e}) = \sigma^2 \mathbf{diag}\,\{k_1,...,k_n\}, \tag{3.3}$$

ktorá popisuje konkrétnu formu heteroskedasticity. Konštanty $k_1, ..., k_n$ musia byť určené ešte pred výpočtom.

V skutočnosti test nezávisí na samotných hodnotách $k, ..., k_n$, ale iba na ich poradí. Alternatívna hypotéza H_1 potom vyjadruje, že rozptyl chýb závisí na niektorej veličine monotónne. Za túto veličinu sa typicky volí jedna z nezávislých premenných alebo vyhladené hodnoty odozvy $\hat{\mathbf{Y}} = \mathbf{Xb}$, kde $\mathbf{b}$ je odhad regresných parametrov β metódou najmenších štvorcov. Túto voľbu použijeme v prípade, keď máme podozrenie, že smerodajná odchýlka rezíduí je priamoúmerná jednej z nezávislých premenných alebo vyhladeným hodnotám odozvy. Potom volíme za konštanty $k_1, ..., k_n$ priamo hodnoty danej nezávislej premennej alebo vyhladené hodnoty odozvy.

Test prebieha nasledovne. Najprv sa spočíta odhad $\mathbf{b}$ parametrov β metódou najmenších štvorcov a rezíduá $\mathbf{u} = \mathbf{Y} - \mathbf{Xb}$. Usporiadame konštanty $k_1, ..., k_n$ od najmenšej po najväčšiu a uvažujeme príslušné pozorovania v rovnakom poradí. Zvyčajne sa ďalej volí prirodzené číslo r tak, aby $r \leq n/2$. Najčastejšou voľbou je $r \doteq n/3$. Potom sa vypočítajú odhady regresných parametrov iba pre model, v ktorom vystupuje prvých r pozorovaní a nezávisle na tom odhady regresných parametrov pre model, v ktorom vystupuje práve posledných r pozorovaní.

Teraz popíšeme testovú štatistiku. Nech SSE_1 značí rezíduálny súčet štvorcov prvého pomocného regresného modelu a SSE_3 druhého pomocného modelu, ktorý bol spočítaný z tretej skupiny pozorovaní a p znázorňuje počet regresorov v lineárnom modeli

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + e_i,\ i = 1, ..., n \tag{3.4}$$

Za predpokladu normality chýb a za platnosti H_0 platí

$$\frac{SSE_1}{\sigma^2} \sim \chi^2_{r-p-1}, \tag{3.5}$$

$$\frac{SSE_3}{\sigma^2} \sim \chi^2_{r-p-1}, \tag{3.6}$$

z čoho vyplýva

$$\frac{SSE_3}{SSE_1} \geq F_{r-p-1,r-p-1}(\alpha), \tag{3.7}$$

kde $F_{r-p-1,r-p-1}(\alpha)$ je kritická hodnota Fisherovho $F_{r-p-1,r-p-1}$ rozdelenia. Pre veľký počet pozorovaní má testová štatistika asymptoticky Fisherovo F-rozdelenie i bez predpokladu normálneho rozdelenia chýb.

Pripomenieme, že rozptyl chýb v lineárnom regresnom modeli (3.4) sa odhaduje pomocou nestranného dohadu ako

$$s^2 = \frac{\mathbf{u}'\mathbf{u}}{n-p-1}. \tag{3.8}$$

Ak označíme pomocou s_1^2 a s_3^2 odhady rozptylov chýb σ^2 spočítané z prvej a tretej časti preusporiadaných dát, čo znamená

$$s_1^2 = \frac{SSE_1}{r-p-1}, \tag{3.9}$$

$$s_3^2 = \frac{SSE_3}{r-p-1}, \tag{3.10}$$

potom sa testová štatistika súčasne rovná podielu $\frac{s_3^2}{s_1^2}$.

Všeobecne je možné pozorovanie rozdeliť do skupín s nerovnakými rozsahmi. Povedzme, že pomocný regresný model pre prvú skupinu dát obsahuje r_1 pozorovaní, zatiaľ čo model pre tretiu skupinu dát obsahuje r_3 pozorovaní. Samozrejme sa požaduje, aby $r_1 > p$, $r_3 > p$, $r_1 + r_3 \leq n$. Definujme SSE_1, SSE_3, s_1^2, s_3^2 podobne ako vyššie. Potom

$$F = \frac{SSE_3}{SSE_1} \cdot \frac{r_1-p-1}{r_3-p-1} = \frac{s_3^2}{s_1^2} \sim F_{r_3-p-1,r_1-p-1}. \tag{3.11}$$

Goldfeldov - Quandtov test je určený pre normálne dáta. Ak nie je splnená normalita chýb, dochádza k veľkému odchýleniu testovej štatistiky (3.11) od F-rozdelenia, čoho príčinou je odchýlenie štatistík (3.5) a (3.5) od χ^2-rozdelenia.

3.2 Breuschov - Paganov test

Ďalším obľúbeným testom je Breuschov - Paganov test, ktorý v roku 1979 navrhli T. S. Breusch a A. R. Pagan [2]. Ich test požaduje, aby heteroskedasticita bola špecifikovaná vo forme, ktorá upresňuje obecnú alternatívu

$$\mathsf{var}(e_i) = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_K Z_{Ki},\ i = 1, ..., n \tag{3.12}$$

pre nejaké veličiny

$$\mathbf{Z}_1 = (Z_{11}, ..., Z_{1n})', ..., \mathbf{Z}_K = (Z_{K1}, ..., Z_{Kn})'. \tag{3.13}$$

Za tie sa najčastejšie volia niektoré alebo všetky z regresorov v pôvodnom lineárnom modeli.

Nulová hypotéza homoskedasticity H_0: $\alpha_1 = ... = \alpha_K = 0$ sa potom testuje proti všeobecnej alternatívnej hypotéze v tvare H_1: H_0 neplatí. Niekedy sa uvažuje všeobecnejšia situácia, keď sa rozptyl chýb modeluje ako nelineárna obdoba vzorca (3.12), teda ako

$$\mathsf{var}(e_i) = h(\alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_K Z_{Ki}),\ i = 1, ..., n, \tag{3.14}$$

kde h je známa nelineárna funkcia. Táto úprava však vedie na zložitejšiu úlohu, ktorá presahuje rozsah tejto práce.

Breusch a Pagan odvodili testovú štatistiku vo forme tzv. skórového testu. Ide o jeden z možných všeobecných postupov pre odvodenie asymptotického testu založeného na vierohodnostnej funkcii za prítomnosti rušivých parametrov. Konkrétne v tomto prípade bol odvodený za predpokladu normálneho rozdelenia náhodných chýb v modeli (3.4).

Breusch a Pagan súčasne ukázali, že ich testová štatistika χ^2 je rovná polovici regresného súčtu štvorcov v pomocnom regresnom modeli

$$\frac{u_i^2}{s^2} = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_K Z_{Ki} + v_i,\ i = 1, ..., n, \tag{3.15}$$

kde $\mathbf{u} = (u_1, ..., u_n)'$ sú rezíduá prislúchajúce odhadu metódou najmenších štvorcov v lineárnom regresnom modeli, s^2 je odhad rozptylu chýb a v_1, ..., v_n sú náhodné chyby. Táto testová štatistika sa už jednoducho spočíta a test zamieta nulovú hypotézu, ak testová štatistika χ^2 prekročí kvantil χ^2_K rozdelenia.

Pre úplnosť uveďme, že Koenker [13] dokázal extrémnu citlivosť Breuschovho - Paganovho testu voči normalite a navrhol drobnú modifikáciu tohto testu, ktorá už taká citlivá nie je. Jeho výsledkami sa však nebudeme v tejto práci podrobnejšie zaoberať.

3.3 Whiteov test

Jedným z najpoužívanejších testov v ekonometrickej praxi je *Whiteov test*. Je špeciálnym prípadom Breuschovho - Paganovho testu pre špeciálnu voľbu pomocného regresného modelu, ktorý popisuje heteroskedasticitu náhodných chýb.

V roku 1980 prišiel Halbert White [15] s nápadom porovnať dva odhady ko-

variančnej matice

$$\frac{1}{n}\sigma^2\mathbf{X}'\mathbf{X}, \tag{3.16}$$

ktoré sú za platnosti homoskedasticity asymptoticky ekvivalentné. Jedným z týchto odhadov bol

$$\frac{1}{n}\hat{\sigma^2}\mathbf{X}'\mathbf{X}, \tag{3.17}$$

kde

$$\hat{\sigma^2} = \frac{1}{n-p-1}\sum_{i=1}^{n} u_i^2 \tag{3.18}$$

a druhým odhadom bol

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 X_i X_i'. \tag{3.19}$$

Ak neplatí hypotéza homoskedasticity, potom budú tieto dva odhady divergovať pre $n \to \infty$. Zisťovanie konvergencie alebo divergencie odhadov je však možné, ako spomína aj Víšek [14], len v prípade dostatočne rozsiahlych súborov.

Tento test je možné založiť na štatistickej významnosti, resp. nevýznamnosti rozdielu druhého (3.19) a prvého (3.17) odhadu kovariančnej matice (3.16). Najjednoduchšia verzia testovej štatistiky pre tento účel je daná vzťahom

$$\frac{1}{n}R^2, \tag{3.20}$$

kde R^2 značí koeficient determinácie (1.9) regresie u^2 na premenné X_{ji}, $i = 1, ..., n$; $j = 1, ..., p$, pričom sú vynechané nadbytočné premenné a pridaná konštanta. Za platnosti nulovej hypotézy má testová štatistika (3.20) asymptotické χ^2-rozdelenie s počtom stupňov voľnosti rovným počtu vysvetľujúcich premenných v práve popísanej regresii, vynímajúc konštantu. Judge [9] dodáva, že ak medzi vysvetľujúcimi premennými nie sú žiadne nadbytočné premenné a zahrnieme medzi ne konštantu, tento počet bude

$$\frac{k(k+1)}{2} - 1. \tag{3.21}$$

Podľa Greena [7] je Whiteov test veľmi všeobecný a k jeho prevedeniu nepotrebujeme robiť žiadne špeciálne predpoklady o povahe heteroskedasticity. Na jednej strane je v tom jeho sila, na strane druhej však vážny nedostatok. Test môže odhaliť heteroskedasticitu, ale namiesto nej môže jednoducho identifikovať iné chyby. Sila tohto testu môže byť veľmi malá, hlavne pre niektoré špeciálne typy heteroskedasticity, keď by bolo vhodnejšie uvažovať alternatívnu hypotézu

v konkrétnejšom tvare.

Nevýhodou tohto testu je jeho nízka výpovedná hodnota, nakoľko nenaznačí, čo robiť v prípade zamietnutia homoskedasticity. Princíp Whiteovho testu, popísaný v knihe Cipra [3], vysvetlíme na konkrétnom modeli. Pritom vysvetlíme a interpretujeme Ciprovo značenie tak, aby odpovedalo nášmu značeniu, ktoré sme zaviedli v prvej kapitole.

Úlohou je previesť test homoskedasticity ako nulovej hypotézy, napr. v modeli

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i,\ i = 1, ..., n. \tag{3.22}$$

Whiteov test v tom prípade vytvorí pomocný model

$$u_i^2 = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i1}^2 + \alpha_4 X_{i2}^2 + \alpha_5 X_{i1} X_{i2} + v_i, \tag{3.23}$$

ktorý je lineárnou regresiou štvorcov rezíduí na konštantu, pôvodné regresory, ich štvorce a ich súčin za predpokladu normálne rozdelenej rezíduálnej zložky v_i. Motívom takéhoto postupu je zistiť, či sa rozptyl pôvodných chýb, ktorý je reprezentovaný ľavou stranou pomocného modelu, systematicky mení v závislosti na všetkých regresoroch pôvodného modelu.

Pred samotným testovaním si zavedieme značenie, ktoré budeme v tomto prípade používať.

- SSE_r označuje *obmedzený rezíduálny súčet štvorcov (restricted error sum of squares)*, pre ktorý platí

$$SSE_r = SST = \sum_{i=1}^{n} (u_i^2 - \bar{u^2}), \tag{3.24}$$

- SSE_u označuje *neobmedzený rezíduálny súčet štvorcov (unrestricted error sum of squares)*, pre ktorý platí

$$SSE_u = SSE = \sum_{i=1}^{n} (u_i^2 - \hat{u_i^2}), \tag{3.25}$$

pričom u_i^2 je ľavá strana v pomocnom modeli (3.23) a $\hat{u_i^2}$ a $\bar{u^2}$ získame z nasledujúcich vzťahov:

$$\bar{u^2} = \frac{1}{n} \sum_{i=1}^{n} u_i^2 \tag{3.26}$$

$$\hat{u_i^2} = \hat{\alpha}_0 + \hat{\alpha}_1 X_{i1} + \hat{\alpha}_2 X_{i2} + \hat{\alpha}_3 X_{i1}^2 + \hat{\alpha}_4 X_{i2}^2 + \hat{\alpha}_5 X_{i1} X_{i2}, \tag{3.27}$$

kde $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5)'$ je vektor odhadov α metódou najmenších štvorcov.

Teraz môžeme pristúpiť k testovaniu. Najprv použijeme χ^2- test, pri ktorom stačí nájsť koeficient determinácie R^2 v pomocnom modeli. Príslušný kritický obor nulovej hypotézy na hladine významnosti α je potom:

$$(n-6) \cdot R^2 \geq \chi^2_{1-\alpha}(5), \tag{3.28}$$

kde k je počet regresorov v pomocnom modeli a m je opäť počet obmedzení.

Alternatívne v modeli (3.23) prevedieme súhrnný *F-test* lineárnych obmedzení

$$H_0 : \alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0, \alpha_5 = 0, \tag{3.29}$$

lebo za platnosti homoskedasticity by to malo platiť. Príslušný kritický obor na hladine významnosti α je potom:

$$\frac{n-6}{5} \cdot \frac{SSE_r - SSE_u}{SSE_u} \geq F_{1-\alpha}(5, n-6), \tag{3.30}$$

lebo v pomocnom modeli bez obmedzení je $k = 6$ a počet lineárnych obmedzení je $m = 5$. SSE_u získame z pomocného modelu (3.23), zatiaľ čo pre SSE_r je nutné pravú stranu pomocného modelu zredukovať na intercept (absolútny člen), teda

$$u_i^2 = \alpha_0 + v_i, \ i = 1, ..., n. \tag{3.31}$$

Ostáva dokázať, že z χ^2-testu je možné odvodiť F-test. Pripomeňme, že koeficient determinácie získame zo vzťahu

$$R^2 = \frac{SST - SSE}{SST}. \tag{3.32}$$

S využitím poznatku, že testová štatistika χ^2-testu je $(n-6) \cdot R^2$, môžeme testovú štatistiku F-testu upraviť nasledovným spôsobom:

$$\begin{aligned} F &= \frac{n-6}{5} \cdot \frac{SST - SSE}{SSE} = (n-6) \cdot \frac{SST - SSE}{SST} \cdot \frac{SST}{5 \cdot SSE} = \\ &= \chi^2 \cdot \frac{SST}{5 \cdot SSE} = \frac{\chi^2}{\frac{SSE}{SST}/5}. \end{aligned} \tag{3.33}$$

Z konštrukcie štatistiky F plynie, že se riadi F-rozdelením, čo platí i vďaka nezávislosti medzi χ^2 a veličinou $\frac{SSE}{SST}$.

Kapitola 4

Testovanie heteroskedasticity v praxi

Táto kapitola sa zaoberá praktickou aplikáciou metód na rozpoznanie heteroskedasticity na konkrétne dáta. V príkladoch budeme nielen testovať heteroskedasticitu, ale aj overovať niektoré predpoklady testov spomínaných v predchádzajúcej kapitole. Pokúsime sa teda o úplné riešenie príkladov nielen z hľadiska heteroskedasticity. Súčasťou tohto riešenia bude aj overovanie normality, na ktoré okrem grafických metód použijeme i Shapirov - Wilkov test normality. Uvedomujeme si však, že testy normality môžu mať všeobecne pre malý počet pozorovaní malú silu.

Heteroskedasticitu budeme testovať na viacerých rôznych súboroch dát, aby sme obsiahli viac možností, ktoré môžu nastať, napr. niektorý z regresorov nie je v modeli významný, dáta nie sú normálne a pod. Na testovanie heteroskedasticity bol použitý štatistický software R, zdrojové kódy k jednotlivým príkladom je možné nájsť v prílohách práce.

4.1 Príklad č. 1: Výdaje vs. príjmy

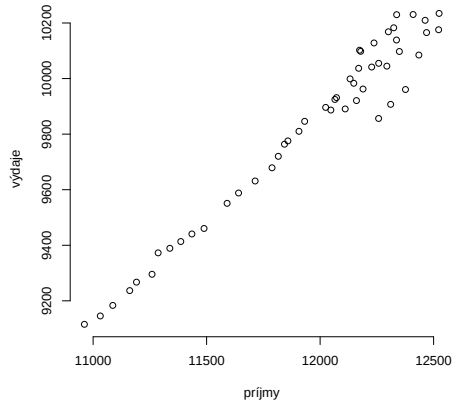
V prvom príklade (príloha č. 4) budeme používať dáta pomenované *Outlays*. Okrem hodnoty osobných výdajov obsahujú i hodnoty osobných príjmov obyvateľstva. Ide o americké dáta [8], v ktorých sú obsiahnuté v mesačnej frekvencii údaje od januára 2006 do februára 2010 uvedené v miliardách amerických dolárov, celkovo máme 50 pozorovaní.

Budeme sa zaoberať lineárnym modelom popisujúcim závislosť výdajov na príjmoch obyvateľstva

$$OUT_i = \beta_0 + \beta_1 INC_i + e_i,\ i = 1, ..., 50, \tag{4.1}$$

kde závislá premenná OUT_i predstavuje hodnoty osobných výdajov a INC_i označuje hodnoty ich príjmov.

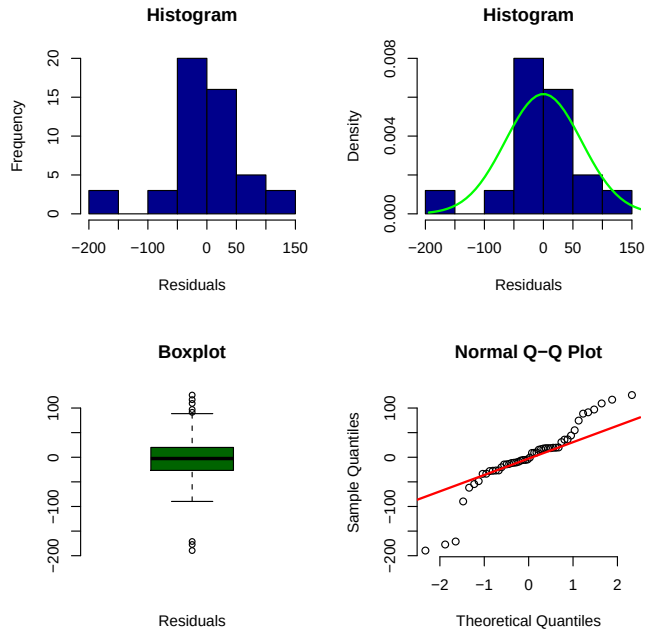
Pred aplikovaním testovacích metód na odhalenie heteroskedasticity overíme niekoľko predpokladov lineárneho regresného modelu. Najprv pomocou grafu závislosti vysvetľovanej premennej OUT_i na vysvetľujúcej premennej INC_i zistíme, či pracujeme s lineárnym modelom. Z grafu č. 4.1 to síce jasné nie je, ale vzhľadom na reálnosť dát to môžeme považovať za akceptovateľné.



Obr. 4.1: Overovanie linearity modelu (4.1) v príklade č. 1

Nasleduje testovanie významnosti regresorov. V tomto prípade máme regresor len jeden a na základe t-testu ho môžeme prehlásiť za významný.

Normalitu chýb $e_i, i = 1, ..., 50$ skúsime najprv odhadnúť z grafov.

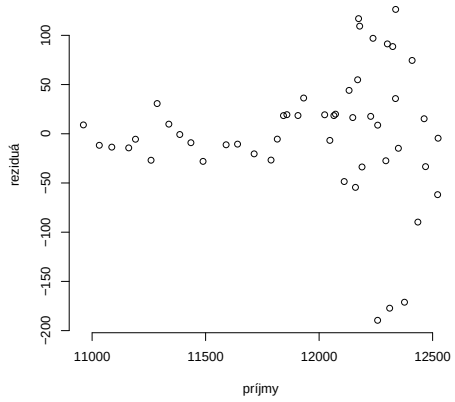


Obr. 4.2: Overovanie normality chýb modelu (4.1) v príklade č. 1

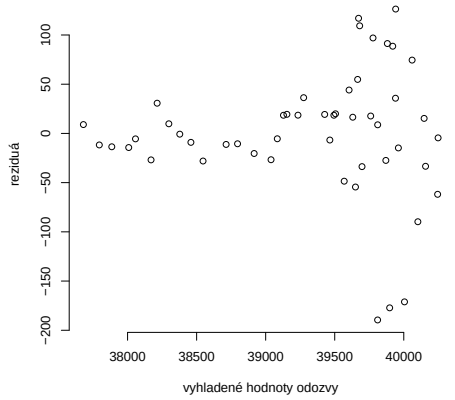
Na základe grafov na obrázku č. 4.2 nemôžeme tvrdiť, že ide o normálne rozdelené chyby. Histogram by mal totiž byť v prípadne normality unimodálny, i keď s malou pravdepodobnosťou môžu i normálne dáta vyzerať inak, boxplot by mal byť symetrický a bez odľahlých pozorovaní a na poslednom z grafov by sa mali body držať okolo priamky. Skúsime aj Shapirov - Wilkov test normality. Keďže p-hodnota je 0,001, na hladine 0,05 zamietame nulovú hypotézu o normalite chýb. Normalita chýb je pri niektorých testoch nutným predpokladom, neponúkajú však možnosť, čo robiť v prípade jej zamietnutia, preto skúsime testovať heteroskedasticitu i na týchto dátach.

Teraz môžeme pristúpiť k samotnému testovaniu heteroskedasticity, najprv na základe grafických metód, neskôr pomocou spomínaných testov heteroskedasticity.

Grafy na obr. č. 4.3 a 4.4 znázorňujú závislosť rezíduí lineárneho modelu (4.1) na regresore *príjem INC* a na vyhladených hodnotách odozvy. V prípade, že by dáta boli heteroskedastické, body by mali ležať rovnomerne okolo nulovej osi, príp. by rezíduá mali rásť s nárastom hodnoty regresoru alebo odozvy. Tvar grafov napovedá o prítomnosti heteroskedasticity. Rezíduá majú nulový priemer, ich absolútna hodnota však rastie s rastom regresoru *príjem INC*, resp. s hodnotami odozvy.



Obr. 4.3: Závislosť rezíduí na príjmoch v modeli (4.1) v príklade č. 1



Obr. 4.4: Závislosť rezíduí na vyhladených hodnotách odozvy v modeli (4.1) v príklade č. 1

Jeden z najpoužívanejších testov heteroskedasticity je Breuschov - Paganov test, ktorý sme popisovali v teoretickej časti tejto práce. Tento test sme skonštruovali v R-ku a použili na lineárny model (4.1). Keďže ide o prvý príklad, konštrukciu Breuschovho - Paganovho testu, ako aj ostatných testov, si vysvetlíme podrobne.

Argumentom funkcie (príloha č. 1) je vektor vysvetľovaných premenných $\mathbf{Y}$ a matica obsahujúca vysvetľujúce premenné, v tomto prípade teda len x označujúcu INC_i, ktorú pomenujeme $\mathbf{X}$. Z takto zadanej matice vieme zistiť počet regresorov p a počet pozorovaní n. K matici pripojíme prvý stĺpec tvorený jednotkovým vektorom a aplikujeme metódu najmenších štvorcov (buď postupnými výpočtami alebo použijeme vstavanú funkciu R-ka *lm*) na výpočet odhadu regresných parametrov a rezíduí. Budeme testovať modifikovaný model (3.15), v našom prípade teda

$$\frac{u_i^2}{s^2} = \alpha_0 + \alpha_1 INC_i + v_i,\ i = 1, ..., 50, \tag{4.2}$$

kde $\mathbf{u}$ je vektor rezíduí modelu (4.1) a s^2 je odhad rozptylu chýb (3.4). Metódu najmenších štvorcov použijeme i na model (4.2), spočítame odhad a priemer novej vysvetľovanej premennej $\frac{u_i^2}{s^2}$ a určíme regresný súčet štvorcov (1.4). Našou testovou štatistikou bude polovica regresného súčtu štvorcov modelu (4.2). Tú porovnáme s kvantilom χ^2-rozdelenia o p stupňoch voľnosti. Spočítame i p-hodnotu testu. Nulovú hypotézu homoskedasticity zamietame, ak hodnota testovej štatistiky je väčšia ako spomínaný kvantil, resp. ak je p-hodnota menšia ako 0,05, homoskedasticitu zamietame na hladine 0,05.

Testová štatistika Breuschovho - Paganovho testu má hodnotu 11,944, kvantil χ^2-rozdelenia $\chi_1^2(0,95)$ je však len 3,841. Hypotézu homoskedasticity teda zamietame na hladine 0,05, čo potvrdzuje aj p-hodnota rovná 0,001.

Teraz na dáta použijeme Goldfeldov - Quandtov test (priloha č. 2). Ako sme sa dozvedeli v podkapitole 3.1, pri tomto teste je potrebné najprv určiť konštantu k, podľa ktorej zoradíme dáta od najmenších po najväčšie hodnoty. V tomto prípade sme za k zvolili hodnoty regresoru INC_i, dáta sme preusporiadali podľa jeho hodnôt a rozdelili sme ich do troch skupín. Na výpočet hodnoty testovej štatistiky potrebujeme len najmenšiu a najväčšiu tretinu dát. Tie v našom prípade obsahovajú 17 pozorovaní. Pre obe časti spočítame odhady regresných parametrov metódou najmenších štvorcov a rezíduálny súčet štvorcov (1.4). Našou testovou štatistikou je teraz podiel rezíduálnych súčtov štvorcov tretej a prvej skupiny dát. Porovnáme ju s kvantilom F-rozdelenia $F_{15,15}(0,95)$, kde r je 17 a p je 3. Dostaneme hodnotu 2,403. Hodnota testovej štatistiky je 37,049 a p-hodnota testu je 4,002 $\times 10^{-9}$. Podľa Goldfeldovho - Quandtovho testu teda homoskedasticitu zamietame.

Ostáva aplikácia posledného z testov popisovaných v tretej kapitole, a to Whiteovho testu (príloha č. 3). Ukážeme si jeho dve varianty - v prvej sa dostaneme k testovej štatistike χ^2-rozdelenia, v druhej budeme testovú štatistiku porovnávať s kvantilom F-rozdelenia.

Argumentom funkcie počítajúcej Whiteov test bude rovnako ako pri predchádzajúcich dvoch testoch vektor $\mathbf{Y}$ reprezentujúci vysvetľovanú premennú a matica $\mathbf{X}$ zložená z vektorov jednotlivých regresorov. Pri oboch variantách postupujeme spočiatku rovnako, teda prevedieme odhad parametrov metódou najmenších štvorcov, na základe ktorého spočítame novú vysvetľovanú premennú. V prípade Whiteovho testu ňou bude druhá mocnina rezíduí $\mathbf{u}$. V pomocnom modeli bude

okrem novej vysvetľovanej premennej vystupovať i modifikovaná matica regresorov. V nej okrem pôvodných použijeme i ich druhé mocniny a súčiny všetkých dvojíc. Dá sa ukázať, že ich celkový počet bude

$$\frac{p \cdot (p+1)}{2} + p. \tag{4.3}$$

V našom prípade bude teda modifikovaný regresný model vyzerať nasledovne

$$u_i^2 = \alpha_0 + \alpha_1 INC_i + \alpha_2 INC_i^2 + v_i,\ i = 1, ..., 50. \tag{4.4}$$

Opäť spočítame metódou najmenších štvorcov odhady regresných parametrov a rezíduá. Ďalší postup sa v rámci dvoch spomínaných variant Whiteovho testu líši. V prvej spočítame celkový súčet štvorcov SST a rezíduálny súčet štvorcov SSE modelu (4.4). Pomocou nich zistíme koeficient determinácie (1.9) tohto modelu, ktorý bude po vynásobení výrazom $(n - 1 - p_{mod})$, kde p_{mod} predstavuje počet regresorov v pomocnom modeli (4.10), našou testovou štatistikou. Hodnota testovej štatistiky je 6,963, kvantil χ^2-rozdelenia $\chi^2_2(0, 95)$ je 5,991 a p-hodnota testu je 0,031. Prvá varianta Whiteovho testu teda homoskedasticitu zamieta.

V druhej variante Whiteovho testu budeme počítať obmedzený a neobmedzený rezíduálny súčet štvorcov. Neobmedzený súčet štvorcov SSE_u je totožný s tým, ktorý sme spočítali v prvej variante Whiteovho testu. Obmedzený súčet štvorcov SSE_r získame z modelu (4.10) po aplikácii lineárnych obmedzení, teda nulových regresných parametrov. Ostane nám teda model

$$u_i^2 = \alpha_0,\ i = 1, ..., 50. \tag{4.5}$$

Štandardným postupom spočítame rezíduálny súčet štvorcov modelu (4.5), a tak získame obmedzený rezíduálny súčet modelu (4.4). Testová štatistika má tentoraz tvar

$$\frac{n - p_{mod} - 1}{p_{mod}} \cdot \frac{SST - SSE}{SSE} \tag{4.6}$$

Druhá alternatíva Whiteovho testu dáva hodnotu testovej štatistiky (4.6) 4,087, kvantil F-rozdelenia $F_{2,47}(0, 95)$ 3,195 a p-hodnotu 0,023. Lineárny regresný model (4.1) teda nemôžeme prehlásiť za homoskedastický.

Výsledky testov zhrnieme do Tabuľky 4.1. Kedže všetky testy zamietajú homoskedasticitu, lineárny regresný model (4.1) môžeme prehlásiť za heteroskedastický.

Tabuľka 4.1: Výsledky testov heteroskedasticity v príklade č. 1

test	hodnota testovej štatistiky	kvantil	p-hodnota
Breuschov - Paganov	11,944	3,841	0,001
Goldfeldov - Quandtov	37,049	2,403	$4{,}002 \times 10^{-9}$
Whiteov (χ^2)	6,963	5,991	0,031
Whiteov (F)	4,087	3,195	0,023

4.2 Príklad č. 2: HDP

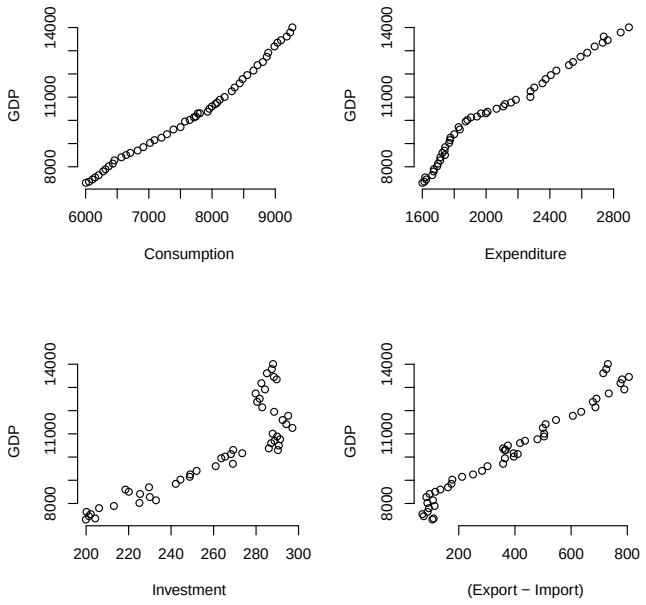
V druhom príklade (príloha č. 5) budeme pracovať s dátami nazvanými *HDP*. Okrem hodnôt hrubého domáceho produktu sú ich súčasťou číselné údaje o spotrebe, vládnych výdajoch, investíciách, importe a exporte. Dáta pochádzajú z USA [8], mapujú obdobie od prvého štvrťroku 1995 do tretieho štvrťroku 2007. Celkovo teda ide o 50 štvrťročných pozorovaní sledovaných veličín, ktoré sú uvedené v miliardách amerických dolárov.

Testovať heteroskedasticitu budeme na lineárnom regresnom modeli popisujúcom závislosť hodnoty HDP na regresoroch

$$\begin{aligned} GDP_i &= \beta_0 + \beta_1 CON_i + \beta_2 EXP_i + \beta_3 INV_i + \beta_4 (IM_i - EX_i) + e_i, \\ &\quad i = 1, ..., 50, \end{aligned} \tag{4.7}$$

kde vysvetľovaná premenná GDP_i predstavuje hodnoty hrubého domáceho produktu, vysvetľujúca premenná CON_i označuje hodnoty spotreby, EXP_i značí hodnoty vládnych výdajov, INV_i predstavuje investície a rozdiel IM_i a EX_i označuje rozdiel medzi importom a exportom.

Ešte pred samotným testovaním homoskedasticity, resp. heteroskedasticity, je dobré overiť linearitu závislosti odozvy na regresoroch. Tú sme zisťovali z obrázku č. 4.5.



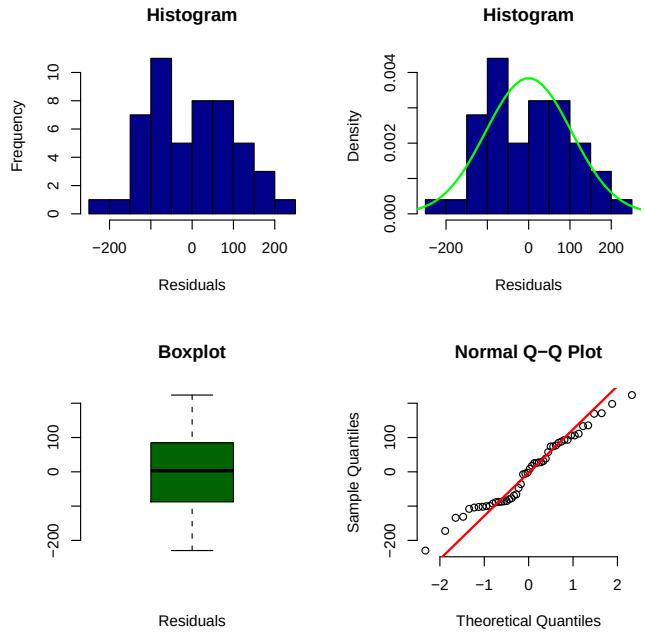
Obr. 4.5: Overovanie linearity regresného modelu (4.7) v príklade č. 2

Keďže ide o reálne dáta a nie o dáta umelo vytvorené, je zrejmé, že vzťah medzi závislou premennou GDP_i a jednotlivými nezávislými premennými nie je úplne lineárny. Našim cieľom je však získať čo najjednoduchší model, ktorý nie je veľmi odchýlený od reálnych dát a lineárna závislosť je zjednodušenie, ktoré nám ľahšie modelovanie umožní.

V regresnom modeli (4.7) vystupujú štyri regresory, nie všetky z nich však musia byť nutne signifikantné. Preto je potrebné t-testmi overiť, či regresné parametre β_0, β_1, β_2, β_3 a β_4 nie sú rovné nule. Na základe výstupu z R-ka nemôžeme vylúčiť nulovosť parametru β_4, teda regresor $(IM_i - EX_i)$ nebudeme považovať za významný a definujeme si modifikovaný model:

$$GDP_i = \beta_0 + \beta_1 CON_i + \beta_2 EXP_i + \beta_3 INV_i + e_i,\ i = 1, ..., 50, \tag{4.8}$$

Ďalším krokom pred samotným testovaním je, rovnako ako v predchádzajúcom príklade, overenie normality chýb e_i, $i = 1, ..., 50$ v modifikovanom modeli (4.8). K tomu opäť použijeme grafy podobne ako pri overovaní linearity modelu (4.7).

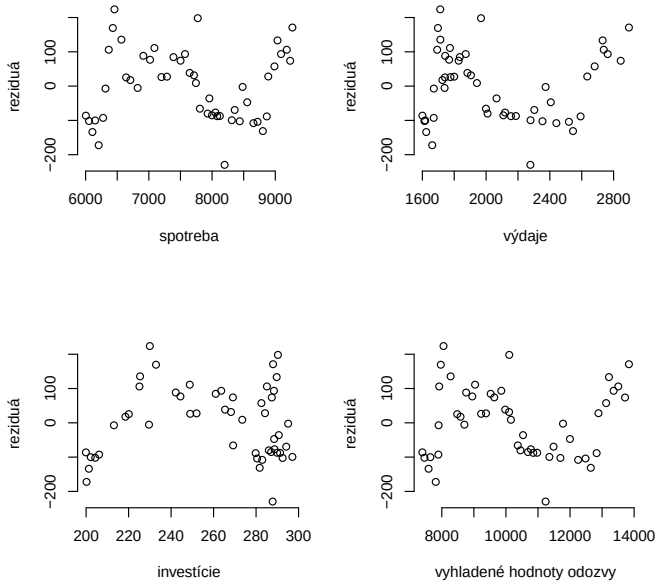


Obr. 4.6: Overovanie normality chýb v modeli (4.8) v príklade č. 2

Z grafov na obrázku č. 4.6 nie je zrejmé, či rozdelenie chýb e_i môžeme považovať za normálne. Podľa krabicového diagramu (boxplot) se chyby zdajú byť symetrické a neobsahujú odľahlé hodnoty. Histogram však nie je unimodálny. Na poslednom grafe by sa mali držať body okolo priamky. To je v našom prípade diskutabilné.

Pri overovaní normality sme použili aj Shapirov - Wilkov test. Na základe p-hodnoty rovnej 0,188 nulovú hypotézu, teda normalitu chýb, na hladine 0,05 nemôžeme zamietnuť.

Teraz už môžeme prejsť k testovaniu heteroskedasticity lineárneho modelu (4.8). Z grafov č. 4.7 popisujúcich závislosť rezíduí na jednotlivých regresoroch a na vyhladených hodnotách odozvy môžeme skúsiť odhadnúť, či sú naše dáta heteroskedastické. Keďže body neležia rovnomerne okolo nulovej osi, nie je zrejmé, či ide o heteroskedastické alebo homoskedastické dáta. Odhadovanie heteroskedasticity z grafov je však veľmi subjektívna metóda testovania.



Obr. 4.7: Overovanie heteroskedasticity v modeli (4.8) v príklade č. 2

Prvým testom, ktorým budeme overovať homoskedasticitu modelu (4.8), je Breuschov - Paganov test. Podobne ako v prvom príklade budeme testovať modifikovaný model (3.15), v tomto prípade teda

$$\frac{u_i^2}{s^2} = \alpha_0 + \alpha_1 CON_i + \alpha_2 EXP_i + \alpha_3 INV_i + v_i,\ i = 1, ..., 50, \tag{4.9}$$

Po aplikácii tohto testu na dáta o HDP sme dostali kvantil $\chi_3^2(0,95)$ o veľkosti 7,815, testovú štatistiku rovnú hodnote 6,568 a p-hodnotu 0,087. Na základe Breuschovho - Paganovho testu teda homoskedasticitu nezamietame.

Teraz na dáta použijeme Goldfeldov - Quandtov test. Za k sme v tomto prípade zvolili hodnoty regresoru CON_i, dáta sme však preusporiadať nemuseli, pretože hodnoty CON_i v tychto dátach rastú s počtom pozorovaní. Podobne by sme mohli test previesť pre iné možné voľby k, napr. pre inú vysvetľujúcu premennú. Budeme pracovať s prvými a poslednými 17 pozorovaniami. Hodnota testovej štatistiky je 5,029 a p-hodnota testu je 0,003, kvantil F-rozdelenia $F_{13,13}(0,95)$ je približne rovný 2,577. Podľa Goldfeldovho - Quandtovho testu teda homoskedasticitu zamietame.

Teraz na dáta aplikujme Whiteov test. Modifikovaný regresný model bude vyzerať nasledovne:

$$
\begin{aligned}
u_i^2 &= \alpha_0 + \alpha_1 CON_i + \alpha_2 EXP_i + \alpha_3 INV_i + \alpha_4 CON_i^2 + \alpha_5 EXP_i^2 \\
&\quad + \alpha_6 INV_i^2 + \alpha_7 CON_i EXP_i + \alpha_8 EXP_i INV_i + \alpha_9 INV_i CON_i + v_i, \\
&\quad i = 1, ..., 50. \qquad (4.10)
\end{aligned}
$$

Hodnota testovej štatistiky je 16,268. Kvantil χ^2-rozdelenia $\chi^2_9(0,95)$ je 16,919 a p-hodnota testu je 0,061. Teda χ^2-varianta Whiteovho testu homoskedasticitu v prvom príklade nezamieta. Hodnota testovej štatistiky v F-variante testu je 3,047, kvantil F-rozdelenia $F_{9,40}(0,95)$ je 2,124 a p-hodnota tohto testu je 0,007. To znamená, že F-varianta Whiteovho testu homoskedasticitu zamieta.

Výsledky, ktoré sme získali testovaním heteroskedasticity v modeli (4.8), sú opäť pre porovnanie uvedené v nasledujúcej tabuľke:

Tabuľka 4.2: Výsledky testov heteroskedasticity v príklade č. 2

test	hodnota testovej štatistiky	kvantil	p-hodnota
Breuschov - Paganov	6,568	7,815	0,087
Goldfeldov - Quandtov	5,029	2,577	0.003
Whiteov (χ^2)	16,268	16,919	0,061
Whiteov (F)	3,047	2,124	0,007

4.3 Príklad č. 3: Výdavky na potraviny

Tretí príklad (príloha č. 6) je venovaný spotrebiteľským výdavkom na potraviny. Budeme pracovať s dátami *Food*, ktorých súčasťou sú okrem hodnôt výdavkov na potraviny aj hodnoty agregovaného disponibilného dôchodku, reálneho a nominálneho cenového indexu potravín a nominálny cenový index pre súhrn osobných výdavkov. Okrem týchto hodnôt sú medzi dátami i zlogaritmované hodnoty prvých troch spomínaných veličín. Spracovávané dáta pochádzajú z USA [5]. Ide o 36 ročných pozorovaní z obdobia rokov 1959 - 1994. Peňažné údaje sú uvedené v miliardách amerických dolárov.

Uvažujme model dopytu po potravinách

$$
F_i = \alpha \cdot (DPI)_i^{\beta_1} \cdot (PRF)_i^{\beta_2} \cdot v_i, \ i = 1, ..., 36 \qquad (4.11)
$$

kde F_i predstavuje výdavky na potraviny, DPI_i je agregovaný disponibilný dôchodok a PRF_i je reálny cenový index potravín. Takýto model (4.11) je analógiou modelov, ktoré se v ekonomickej literatúre obvykle používajú pre modelovanie

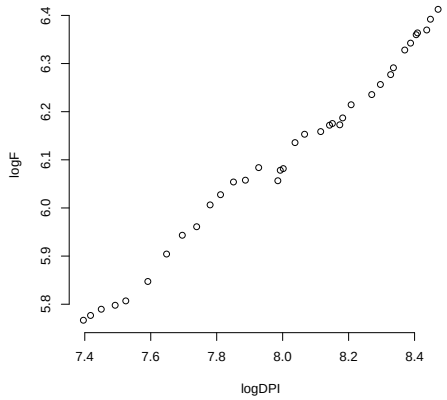
závislosti výdajov na dalších ekonomických veličinách [7]. Reálny cenový index potravín sa vypočíta zo vzťahu

$$PRF = 100 \cdot \frac{PF}{PTPE}, \tag{4.12}$$

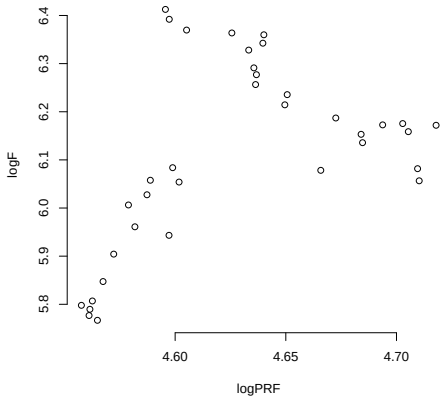
kde PF je nominálny cenový index potravín a $PTPE$ označuje nominálny cenový index pre súhrn osobných výdavkov. Zlogaritmovaním pôvodného modelu a nahradením $log(\alpha) := \beta_0$ a $log(v)_i := e_i$ získame model, s ktorým budeme pracovať

$$\log F_i = \beta_0 + \beta_1 log(DPI)_i + \beta_2 log(PRF)_i + e_i,\ i = 1, ..., 36 \tag{4.13}$$

Pri overovaní predpokladov lineárneho modelu (4.13) postupujeme rovnako ako v predchádzajúcich dvoch príkladoch. Čo sa týka linearity, z grafu 4.8 je vidieť, že pri vysvetľujúcej premennej logDPI sa s ňou počítať môže, pri druhej vysvetľujúcej premennej, ako sa môžeme presvedčiť pohľadom na graf č. 4.9, sa však o linearite hovoriť nedá.



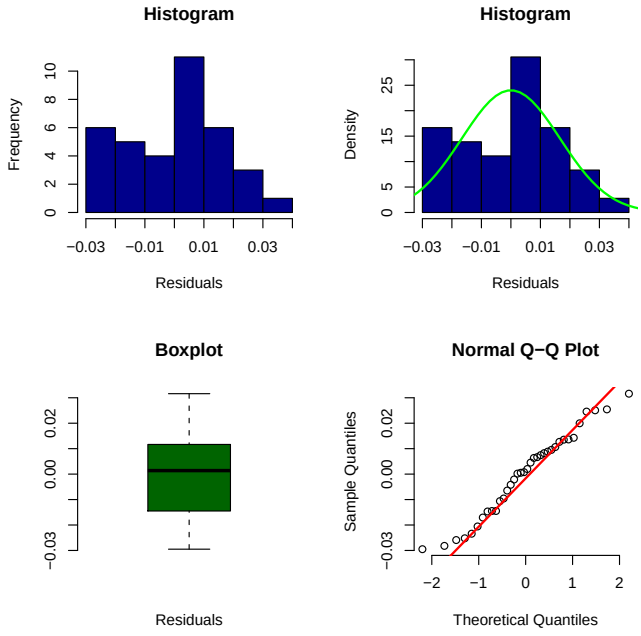
Obr. 4.8: Závislosť logDPI na logF v modeli (4.13) v príklade č. 3



Obr. 4.9: Závislosť logPRF na logF v modeli (4.13) v príklade č. 3

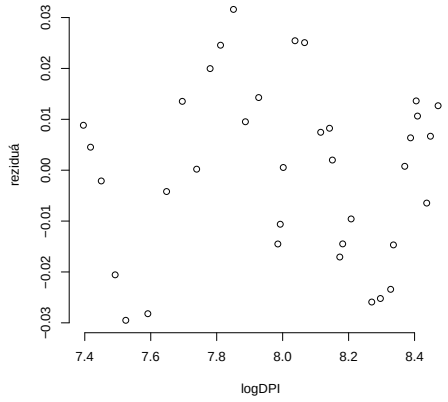
Teraz sa pozrieme na významnosť regresorov $logDPI$ a $logPRF$. Na základe t-testov môžeme oba prehlásiť za významné, model teda nemusíme modifikovať.

Ostáva overiť normalitu rezíduí modelu (4.13). Najprv sa pozrieme na obrázok č. 4.10. Podľa histogramov nie je jasné, či môžeme normalitu zamietnuť, podľa boxplot grafu a Q-Q grafu by rezíduá mohli byť normálne. Použijeme i Shapirov - Wilkov test. Keďže p-hodnota je 0,363, na hladine 0,05 normalitu nezamietame.

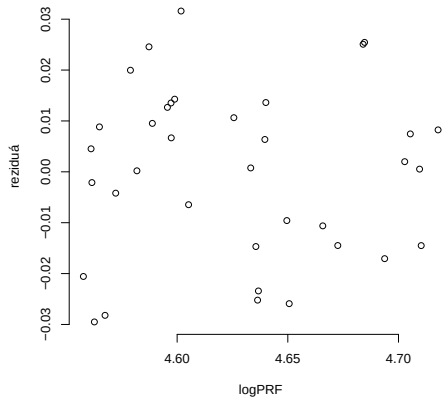


Obr. 4.10: Overovanie normality rezíduí modelu (4.13) v príklade č. 3

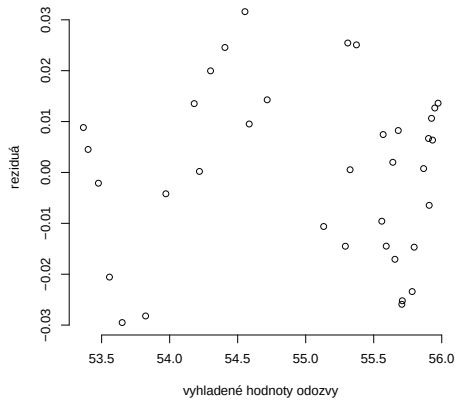
Môžeme pristúpiť k heteroskedasticite. Grafy závislosti rezíduí na regresoroch a na hodnotách odozvy č. 4.11, 4.12 a 4.13 napovedajú o tom, že heteroskedasticita v modeli (4.13) nie je prítomná.



Obr. 4.11: Závislosť rezíduí na logDPI v modeli (4.13) v príklade č. 3



Obr. 4.12: Závislosť rezíduí na logPRF v modeli (4.13) v príklade č. 3



Obr. 4.13: Závislosť rezíduí na vyhladených hodnotách odozvy v modeli (4.13) v príklade č. 3

Rovnaký výsledok ako grafická metóda nám dáva i aplikácia Breuschovho - Paganovho testu na lineárny model (4.13). Testová štatistika má hodnotu 0,426, kvantil $\chi^2_2(0,95) = 5{,}991$ a p-hodnota je rovná 0,808, teda test homoskedasticitu nezamieta.

Pred použitím Goldfeld - Quandtovho testu sme dáta usporiadali podľa veľkosti regresoru logPRF. Keďže dáta majú 36 pozorovaní, rozdelili sme ich do troch skupín po 12. Podľa tohto testu model (4.13) nie je heteroskedastický. Testová štatistika je totiž rovná 0,616 a kvantil F-rozdelenia $F_{9,9}(0,95) = 3{,}179$. P-hodnota testu je 0,759.

Ostáva použitie Whiteovho testu. V jeho prvej alternatíve je testová štatistika rovná 4,759, kvantil $\chi^2_5(0,95) = 11.071$ a p-hodnota testu je 0,446. Na hladine 0,05 teda Whiteov test homoskedasticitu nezamieta. K rovnakému záveru prídeme aj použitím druhej verzie Whiteovho testu. V tom prípade výjde testová štatistika rovná hodnote 1,131, kvantil $F_{5,30}(0,95) = 2{,}534$ a p-hodnota je 0,365. Testy heteroskedasticity, ktoré sme použili, teda zhodne nevedú k zamietnutiu nulovej hypotézy homoskedasticity.

Výsledky všetkých testov, ktoré sme aplikovali na lineárny model (4.13), sú prehľadne uvedené i v nasledujúcej tabuľke:

Tabuľka 4.3: Výsledky testov heteroskedasticity v príklade č. 3

test	hodnota testovej štatistiky	kvantil	p-hodnota
Breuschov - Paganov	0,426	5,991	0,808
Goldfeldov - Quandtov	0,616	3,179	0,759
Whiteov (χ^2)	4,759	11.071	0,446
Whiteov (F)	1,131	2,534	0,365

Záver

V tejto práci sme sa venovali konkrétnemu problému v lineárnej regresii, a to heteroskedasticite a jej testovaniu. Tento problém býva často podceňovaný, no jeho prehliadnutím môžeme pri práci s lineárnymi regresnými modelmi dospieť k mylným výsledkom, čo môže mať závažné následky.

V ekonometrickej literatúre je heteroskedasticite venovaný značný priestor, v štatistickej však už menej. Práve z ekonometrických zdrojov pochádza väčšina teoretických informácií a poznatkov, ktoré sme použili v tejto práci. Od teoretických základov týkajúcich sa lineárnej regresie sme prešli k heteroskedasticite, jej dôsledkom, riešeniu a testovaniu. Popísali sme najpoužívanejšie metódy jej odhaľovania a najznámejšie testy heteroskedasticity. Následne sme tieto metódy a postupy aplikovali prakticky na konkrétne dáta.

Testovaním heteroskedasticity rôznymi testami - Goldfeldovým - Quandtovým, Breuschovým - Paganovým a Whiteovým sme zistili, že nie všetky z nich nám musia vždy dávať rovnaké výsledky. Podľa niektorého môžeme pracovať s homoskedastickými dátami, podľa iného sa prítomnosť homoskedasticity zamieta. Tieto testy sa nedajú používať všeobecne na všetky lineárne regresné modely. Sú limitované určitými predpokladmi, o ktorých sme v tejto práci diskutovali, a taktiež zaťažené určitou mierou subjektivity. Nie je k dispozícii presné pravidlo, ktorý z testov zvoliť pre analýzu konkrétnych dát. Je preto potrebné vopred zvažiť, ktorý test je vhodné v našom prípade použiť, aby sme sa vyhli nepresným záverom.

Literatúra

[1] ANDĚL, J. *Základy matematické statistiky*. Praha: Matfyzpress, 2007. ISBN: 80-7378-001-1.

[2] BREUSCH, T. S.; PAGAN, A. R. *A simple test for heteroscedasticity and random coeficient variation*. In: Econometrica, 1979, roč. 47, č. 5, s. 1287 - 1294.

[3] CIPRA, T. *Finanční ekonometrie*. Praha: Ekopress, 2008. ISBN: 978-80-86929-43-9.

[4] CRAGG, J. G. *More efficient estimation in the presence of heteroscedasticity of unknown form*. In: Econometrica, 1983, roč. 51, č. 3, s. 751 - 763.

[5] Food. [online]. [citované 15. 02. 2010]. Dostupné z <http://www.census.gov/>.

[6] GOLDFELD, S. M.; QUANDT, R. E. *Some tests for homoscedasticity*. In: Journal of the American Statistical Association, 1965, roč. 60, č. 310, s. 539 - 547.

[7] GREENE, W. H. *Econometric analysis*. New Jersey: Prentice Hall, 1993. ISBN: 0-13-373549-4.

[8] Gross Domestic Product (GDP) and Components. [online]. [citované 2. 04. 2011]. Dostupné z <http://research.stlouisfed.org/fred2/categories/18>.

[9] JUDGE, G. G.; GRIFFITHS, W. E.; HILL R. C.; LUTKEPOHL, H.; LEE, T. Ch. *The Theory and Practice of Econometrics*. New York: J. Wiley and Sons, 1985. ISBN: 0-471-89530-X.

[10] KALINA, J. *Least weighted squares in econometric applications*. In: Journal of Applied Mathematics, Statistics and Informatics, 2009, roč. 5, č. 2, s. 115 - 125.

[11] KALINA, J. *On multivariate methods in robust econometrics*. In: Prague economic papers, 2011, prijaté, v tlači.

[12] KMENTA, J. *Elements of Econometrics*. New York: The Macmillan Company, 1986. ISBN: 0-02-946252-5.

[13] KOENKER, R. *A note on studentizing a test for heteroskedasticity*. In: Journal of Econometrics, 1981, č. 17, s. 107 - 112.

[14] VÍŠEK, J. A *Ekonometrie I.*. Praha: Karolinum, 1997. ISBN: 80-7184-483-7.

[15] WHITE, H. *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. In: Econometrica, 1980, roč. 48, č. 4, s. 817 - 838.

[16] ZVÁRA, K. *Regrese*. Praha: Matfyzpress, 2008. ISBN: 978-80-7378-041-8.

Zoznam obrázkov

Zoznam tabuliek

Zoznam použitých skratiek

SSE reziduálny súčet štvorcov (error sum of squares)

SSR regresný súčet štvorcov (regression sum of squares)

SST celkový súčet štvorcov (total sum of squares)

SSE$_r$ obmedzený reziduálny súčet štvorcov (restricted error sum of squares)

SSE$_u$ neobmedzený reziduálny súčet štvorcov (unrestricted error sum of squares)

Zoznam príloh

Na priloženom CD sa nachádzajú nasledujúce prílohy:

Príloha č. 1: Breuschov - Paganov test (zdrojový kód v R - .pdf)

Príloha č. 2: Goldfeldov - Quandtov test (zdrojový kód v R - .pdf)

Príloha č. 3: Whiteov test (zdrojový kód v R - .pdf)

Príloha č. 4: Príklad č. 1 - Výdaje vs. príjmy (zdrojový kód v R - .pdf)

Príloha č. 5: Príklad č. 2 - HDP (zdrojový kód v R - .pdf)

Príloha č. 6: Príklad č. 3 - Výdavky na potraviny (zdrojový kód v R - .pdf)

Príloha č. 7: Breuschov - Paganov test (zdrojový kód v R - .txt)

Príloha č. 8: Goldfeldov - Quandtov test (zdrojový kód v R - .txt)

Príloha č. 9: Whiteov test (zdrojový kód v R - .txt)

Príloha č. 10: Príklad č. 1 - Výdaje vs. príjmy (zdrojový kód v R - .txt)

Príloha č. 11: Príklad č. 2 - HDP (zdrojový kód v R - .txt)

Príloha č. 12: Príklad č. 3 - Výdavky na potraviny (zdrojový kód v R - .txt)

Príloha č. 13: outlays.txt - dáta k príkladu č. 1

Príloha č. 14: HDP.txt - dáta k príkladu č. 2

Príloha č. 15: FOOD.txt - dáta k príkladu č. 3

Príloha č. 16: SpakovaMariaBP.pdf - elektronická verzia tejto práce